

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-108894

(43)Date of publication of application : 12.04.2002

(51)Int.Cl.

G06F 17/30

G06F 17/21

G06F 17/28

(21)Application number : 2000-293597

(71)Applicant : RICOH CO LTD

(22)Date of filing : 27.09.2000

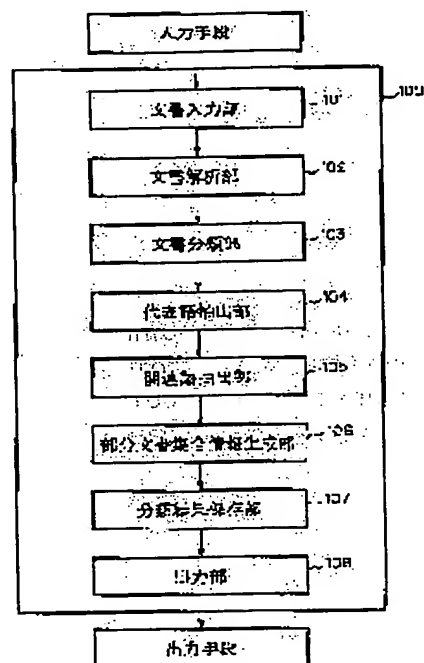
(72)Inventor : KENMOCHI EIJI

(54) DEVICE AND METHOD FOR SORTING DOCUMENT AND RECORDING MEDIUM FOR EXECUTING THE METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a document sorting device which can supply information effective for analyzing a partial document set and also can extract many analysis information from a document set.

SOLUTION: A document analysis part 102 extracts word information from a document that is inputted from a document input part 101 and a document sorting part 103 sorts the document into a partial document set on the basis of the word information. A keyword extraction part 104 extracts a keyword set from the partial document set and a relative word extraction part 105 extracts a relative word set of the partial document set by using a relative word dictionary. A partial document set information generation part 106 generates information on each partial document set and the relative information on these partial document sets on the basis of information on the relative word set, the keyword set and the document set of partial document set. Then a sorting result preservation part 107 preserves the sorting result of the part 103 and information generated at the part 106.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号  
特開2002-108894  
(P2002-108894A)

(43) 公開日 平成14年4月12日 (2002.4.12)

| (51) Int.Cl. <sup>7</sup> | 識別記号  | F I           | テームコード* (参考)      |
|---------------------------|-------|---------------|-------------------|
| G 0 6 F 17/30             | 2 1 0 | G 0 6 F 17/30 | 2 1 0 D 5 B 0 0 9 |
|                           | 1 7 0 |               | 1 7 0 A 5 B 0 7 5 |
| 17/21                     | 5 5 0 | 17/21         | 5 5 0 A 5 B 0 9 1 |
| 17/28                     |       | 17/28         | Z                 |

審査請求 未請求 請求項の数39 O L (全 34 頁)

(21) 出願番号 特願2000-293597(P2000-293597)

(22) 出願日 平成12年9月27日 (2000.9.27)

(71) 出願人 000006747

株式会社リコー

東京都大田区中馬込1丁目3番6号

(72) 発明者 剣持 栄治

東京都大田区中馬込1丁目3番6号 株式  
会社リコー内

(74) 代理人 100079843

弁理士 高野 明近 (外2名)

Fターム(参考) 5B009 MC04 QA01

5B075 ND03 NK35 NR12 QP03

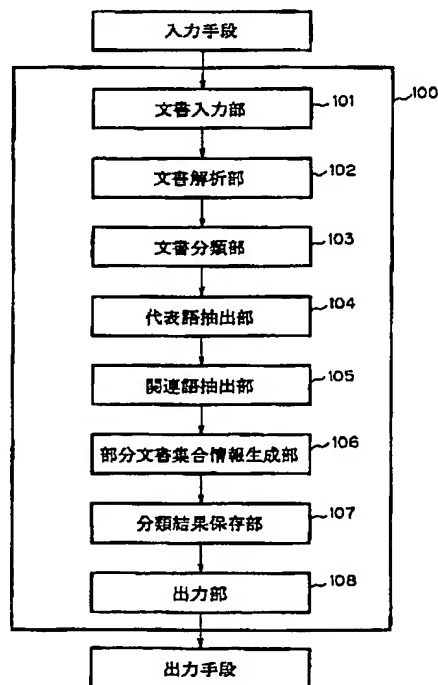
5B091 AB12 AB13 AB17 CA02

(54) 【発明の名称】 文書分類装置、文書分類方法及び該方法を実行するための記録媒体

(57) 【要約】

【課題】 部分文書集合の分析に有効な情報を提供し、文書集合からより多くの分析情報を抽出しうる文書分類装置を提供する。

【解決手段】 文書解析部102は、文書入力部101から入力した文書から単語情報を抽出し、文書分類部103がこれをもとに文書を部分文書集合に分類する。代表語抽出部104は、部分文書集合から代表語セットを抽出し、関連語抽出部105が、関連語辞書を用いて部分文書集合の関連語セットを抽出する。部分文書集合情報生成部106は、関連語セット及び代表語セットと部分文書集合の文書集合に関する情報をもとに個々の部分文書集合及びこれらの間の関連情報を生成し、分類結果保存部107が文書分類部103の分類結果と部分集合情報生成部106で生成された情報とをあわせて保存する。



## 【特許請求の範囲】

【請求項 1】 文書集合をその内容に従って分類する文書分類装置であって、複数の文書を入力する文書入力部と、該文書入力部にて入力された各文書から該各文書を構成する単語情報を抽出する文書解析部と、該文書解析部にて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類部と、該文書分類部にて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出部と、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出部にて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出部と、該関連語抽出部にて抽出した関連語セットと前記代表語抽出部にて抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成部と、前記文書分類部での分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存部とを含むことを特徴とする文書分類装置。

【請求項 2】 請求項 1 に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴とする文書分類装置。

【請求項 3】 請求項 1 に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴とする文書分類装置。

【請求項 4】 請求項 1 に記載の文書分類装置において、前記関連語抽出部にて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴とする文書分類装置。

【請求項 5】 文書の内容に従って文書の分類を行う文書分類装置であって、文書データを入力する文書入力部と、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類部と、該文書分類部で生成された分類結果を記憶する分類結果記憶部と、前記分類数決定部から前記分類結果記憶部までの処理を繰り返し行うか否かの判定をおこなう繰り返し判定部と、前記文書ベクトルデータ記憶部と前記分類結果記憶部に記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出部と、該部分文書集合間関係算出

ル空間生成部と、該文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類部とを含み、前記文書解析部にて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴とする文書分類装置。

【請求項 6】 請求項 5 に記載の文書分類装置において、前記文書分類部において統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴とする文書分類装置。

【請求項 7】 請求項 5 または 6 に記載の文書分類装置において、前記文書解析部において品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うことを特徴とする文書分類装置。

【請求項 8】 請求項 5 ないし 7 のいずれか 1 に記載の文書分類装置において、前記文書解析部において特定の品詞の単語が出現するまで単語の結合を続けることを特徴とする文書分類装置。

【請求項 9】 請求項 5 ないし 8 のいずれか 1 に記載の文書分類装置において、前記文書解析部において品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類部では削除した単語の情報をを用いないことを特徴とする文書分類装置。

【請求項 10】 文書の内容に従って文書データ集合を分類する文書分類装置であって、文書データ集合を入力する文書入力部と、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析結果を記憶する文書解析結果記憶部と、前記文書解析部にて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶部と、指定される条件から文書データ集合の分類数を決定する分類数決定部と、前記文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類部と、該文書分類部で生成された分類結果を記憶する分類結果記憶部と、前記分類数決定部から前記分類結果記憶部までの処理を繰り返し行うか否かの判定をおこなう繰り返し判定部と、前記文書ベクトルデータ記憶部と前記分類結果記憶部に記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出部と、該部分文書集合間関係算出

部にて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶部とを含むことを特徴とする文書分類装置。

【請求項 11】 請求項 10 に記載の文書分類装置において、前記文書分類部にて用いられる統計手法が非階層クラスタリング手法であることを特徴とする文書分類装置。

【請求項 12】 請求項 10 または 11 に記載の文書分類装置において、前記部分文書集合間関係算出部にて算出される関係が、類似関係と包含関係であることを特徴とする文書分類装置。

【請求項 13】 請求項 12 に記載の文書分類装置において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴とする文書分類装置。

【請求項 14】 文書集合をその内容に従って分類する文書分類方法であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含むことを特徴とする文書分類方法。

【請求項 15】 請求項 14 に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴とする文書分類方法。

【請求項 16】 請求項 14 に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴とする文書分類方

法。

【請求項 17】 請求項 14 に記載の文書分類方法において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴とする文書分類方法。

【請求項 18】 文書の内容に従って文書の分類を行う文書分類方法であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴とする文書分類装置。

【請求項 19】 請求項 18 に記載の文書分類方法において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴とする文書分類方法。

【請求項 20】 請求項 18 または 19 に記載の文書分類方法において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行うことを特徴とする文書分類方法。

【請求項 21】 請求項 18 ないし 20 のいずれか 1 に記載の文書分類方法において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続けることを特徴とする文書分類方法。

【請求項 22】 請求項 18 ないし 21 のいずれか 1 に記載の文書分類方法において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いないことを特徴とする文書分類方法。

【請求項 23】 文書の内容に従って文書データ集合を分類する文書分類方法であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素

解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含むことを特徴とする文書分類方法。

【請求項 24】 請求項 23 に記載の文書分類方法において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法であることを特徴とする文書分類方法。

【請求項 25】 請求項 23 または 24 に記載の文書分類方法において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係であることを特徴とする文書分類方法。

【請求項 26】 請求項 25 に記載の文書分類方法において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴とする文書分類方法。

【請求項 27】 文書集合をその内容に従って分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれに

ついて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 28】 請求項 27 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせである文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 29】 請求項 27 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 30】 請求項 27 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 31】 文書の内容に従って文書の分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現する

ための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換える文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 2】 請求項 3 1 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 3】 請求項 3 1 または 3 2 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 4】 請求項 3 1 ないし 3 3 のいずれか 1 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続ける文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 5】 請求項 3 1 ないし 3 4 のいずれか 1 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報をを用いない文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 6】 文書の内容に従って文書データ集合を分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書デ

タの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 7】 請求項 3 6 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 8】 請求項 3 6 または 3 7 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 3 9】 請求項 3 8 に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出される文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書分類装置、文書分類方法及び該方法を実行するための記録媒体に関し、情報分類、情報分析、情報検索等に応用可能な文書分類技術に関する。

【0002】

【従来の技術】インターネット等の普及により大量の文書情報へのアクセスが可能になり、収集した大量の文書情報を意味のあるグループに分類し、文書集合の構造を

10

20

30

40

50

把握するなどの知的作業が行われ始めている。大量な文書集合を分析する場合、まず文書集合をいくつかの話題で分類し、得られた部分文書集合（ある基準で集められた複数の文書）を単位としてさまざまな作業を行うことで、分析作業を効率的に行うことができるものと考えられる。大量の文書情報をユーザが手動で分類する場合、人的／時間的コストが膨大なものになるため、文書集合を文書の内容により自動分類できる装置が望まれている。

【0003】従来、膨大な文書集合からの質の高い分類結果を得るための発明が広く行われている。例えば、特開平7-36897号公報に記載の発明は、分類対象文書集合に含まれる単語を特徴量とする文書特徴ベクトルを用い、その文書特徴ベクトルに対してクラスタリング手法を適用して分類を行うものである。上記の発明ではユーザの意図を反映した分類を行うためにクラスタリングの初期重心ベクトルをユーザが指定することも示唆している。

【0004】また、特開平11-296552号公報に記載の発明は、単語の多義性/同義性を考慮するために文書間の内積行列に特異値分解を適用することにより文書間の単語の共起性を基に潜在的意味空間を生成して、文書と単語を潜在的意味空間に射影し、その潜在的意味空間においてクラスタリング手法などを用いて文書分類を行うものである。このように膨大な文書集合からの質の高い分類結果を得るための発明は種々提案されているが、文書集合の分析を行うためには文書集合を分類するだけでは不十分であり、生成された部分文書集合からどのように有効な情報を抽出するかということも重要な問題であるが、この点についての発明はあまり見られない。

【0005】また、形態素解析などの自然言語処理を用いて文書からそれらを構成する単語を抽出することにより文書を単語頻度のベクトル（文書特徴ベクトル）として空間表現することが可能となるが、これは文書ベクトル空間モデルと呼ばれ、広く用いられている。上述した特開平7-36897号公報の発明は、このような文書ベクトル空間において、クラスタリング手法を適用することにより文書分類を行うものである。

【0006】このように文書ベクトル空間で統計的手法を用いて文書分類処理や文書検索処理等を行う場合、文書ベクトル空間が異なれば得られる結果の質も変わると考えられるので、如何にして良い文書ベクトル空間を生成するかが高品位な処理結果を得るためには重要な問題となる。

【0007】前述したように、通常文書ベクトル空間の各軸は分類対象文書データに形態素解析を適用した結果抽出される単語をもとに構成されるため、例えば、特開平11-110408号公報や特開平11-259487号公報に代表される発明は、検索問い合わせ語や検索対象文書に対し、形態素解析を適用し、その結果抽出さ

れる単語から適切な条件のもとに複合語を生成し、これらの複合語の情報も前記文書ベクトル空間の生成に用いることで、文書ベクトル空間上で行う文書検索の精度の向上を目的としている。従って、文書ベクトル空間で文書分類処理を行う場合においても、複合語を考慮して文書ベクトル空間を生成することで高品位な分類結果を得ることが期待される。

【0008】ところで、上記先願を含め、通常複合語を考慮する場合は、品詞が名詞もしくはそれに類するものが対象とされているが、名詞だけでなく他の結合可能な品詞も適切に結合させることで、より高品位な文書ベクトル空間を構成することが可能になると考えられる。すなわち、先願の発明等ではあまり扱われることのなかった、接頭詞、接尾詞、助数詞、及びそれらに類する品詞を有する単語について、適切な基準でそれらの前後の単語と結合することで生成される単語と置き換えるとともに、品詞も適切なものに置き換えることを考える。

【0009】例えば、“イタリア製の車”という文字列に対して形態素解析を適用し、“イタリア【普通名詞】、製【接頭詞】、の【格助詞】、車”という結果が得られた場合、接頭詞である“製”という単語に着目し、これをこの直前に抽出されている“イタリア”という普通名詞と結合し、“イタリア製”という単語を生成し、これを普通名詞の品詞を有する単語として、“製”という単語と置き換える。そして、この文字列に加え、“イタリアの特色”、“イタリア製の皿”という文字列で構成するベクトル空間を生成することを考えてみる。

【0010】名詞だけで空間を生成することを考えた場合、前記の結合・置き換え処理を行わない場合、ベクトル空間を構成する単語は、“イタリア、車、特色、皿”であり、前記文字列は、単語の出現頻度を座標値と考えた場合、 $(1, 0, 1, 0)$ 、 $(1, 0, 0, 0)$ 、 $(1, 0, 0, 1)$ となる。この場合、前記3つの文字列の相互の類似度をベクトル間の内積で計算すると、前記3つの文字列の相互の類似度は同じものとなる。一方、前記の結合・置き換え処理を行った後、名詞で空間を生成すると、ベクトル空間を構成する単語は、“イタリア、イタリア製、車、特色、皿”となる。同様に、前記3文書のこの空間でのベクトルは、 $(1, 1, 0, 1, 0)$ 、 $(1, 0, 0, 0, 0)$ 、 $(1, 1, 0, 0, 1)$ となる。この場合、前記3つの文字列の相互の類似度には、差異が生じ、最初の文字列と最後の文字列が2番目の文字列より高い類似度を持つことになる。すなわち、この結合・置き換え処理によりベクトル空間により限定化された意味を測る特徴次元を加えることができ、これによりこのベクトル空間で行う文書分類等の質も向上するものと考えられる。

【0011】また、“2000年の目標”という文字列に対し形態素解析を適用し、“2000【数詞】、年【助数詞】、の【格助詞】、目標【普通名詞】”のよう

な結果が得られているとする。このとき、助数詞である“年”という単語に着目し、これをこの直前に抽出されている“2000”という数詞と結合し、“2000年”という単語を生成し、これを普通名詞の品詞を有する単語として、“年”という単語と置き換え、かつ“2000”という数詞を削除する。これにより、非常に漠然とした意味しか有していない助数詞である“年”や“2000”という単語にかえて、“2000年”というより意味的に限定された、それゆえ変数としてはより重要な単語をもとにして文書ベクトル空間が構成可能になることが期待される。

【0012】また、上述のようにインターネット等の普及により大量の文書データへのアクセスが可能になり、その結果として興味のある情報が記述されている文書データを簡単にかつ大量に収集できるようになったが、しかしその一方で、収集した文書データが大量であるがために、それら文書データから有効な情報を読み取る作業は非常に困難なものになってしまっている。このため、大量の文書データから自動もしくは半自動で有効な情報を簡単に抽出することを目的として、文書検索や文書自動分類に関する研究・開発が盛んに行われている。特に、文書分類手法は、生成される複数の部分文書データ集合個々を文書データに含まれる複数の話題を示すものと考え、文書データ全体の構造を把握する手法として非常に有効なものである。

【0013】上述のような目的のために開発された手法の代表的なものに、Scatter/Gather法(D. Cutting et al., Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections., Proc. ACM SIGIR '92)がある。Scatter/Gather法では、文書データ集合の話題を代表文書と代表単語により表現するとともに、話題が不明瞭な文書集合に対して逐次クラスタリングを適用し、複数の部分文書データ集合に分割していくことで文書集合に含まれる様々な話題を理解していく。文書集合の構造を理解するためには、文書集合に含まれる部分文書集合個々を理解することはもちろん必要であるが、加えて部分文書集合間の関係に関する情報も必要であると考えられる。しかしながら、Scatter/Gather法では個々の部分文書集合に関する情報しか提示されていないため、Scatter/Gather法のみでは文書集合の構造を把握することは困難であると考えられる。

【0014】また、一般的に文書分類手法においては、生成する部分文書集合の数が実行前に必要であるが、最適な部分文書集合の数を予測することは極めて困難である。しかも、一方で生成する部分文書集合の数が異なれば生成される部分文書集合の構造も変化してしまう。このため、必要な情報を得るためには生成する部分文書集合の数をかえながら、繰り返し文書分類を行わなければならない。Scatter/Gather法はこの点についても一つの解決法を提示しており、ユーザがより詳細な構造を知

りたいと考える部分文書集合のみに対し逐次クラスタリングを適用し、あらたな部分文書集合を生成し、それらを詳細に分析することで所望の情報を得ることができるとともに、この行為により文書集合全体の構造を理解することも容易になっていると考えられる。

【0015】すなわち、ユーザが行いたいことは文書集合の構造の把握であり、部分文書集合を生成するという行為は本来ユーザが行う必要がないものと考えられる。そして、ユーザが、事前に文書集合から様々な数の部分文書集合を生成し、生成された多数の部分文書集合間の関係を算出しておくことで、ユーザは始めから構造の把握を行う作業に集中できると考えられる。しかしながら、前述の通りScatter/Gather法は部分文書間の関連に関しては考慮されていない。

【0016】

【発明が解決しようとする課題】本発明の請求項1～4、14～17、及び27～30の発明では、文書分類を行うとともに、単語の関連語情報を基に生成された部分文書集合個々及びそれらの関連情報をさらに生成することで部分文書集合の分析に有効な情報を提供することを目的とする。さらに、関連語として反対語に着目することで、各部分文書集合の代表語セットの反対語を含む部分文書集合が、生成された部分文書集合にはない場合、反対語を含むあらたな部分文書集合を生成することで、文書集合からより多くの分析情報を抽出しうる文書分類装置を提供することを目的とする。

【0017】従って請求項1、14及び27の発明は、生成された部分文書集合それぞれの代表語セットを抽出し、さらにそれら代表語それぞれについて関連語を求め、これらの情報をもとに各部分文書集合および部分文書集合間の関連情報を生成することで、部分文書集合の分析に有効な情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【0018】請求項2、15及び28の発明は、関連語として同義語、類義語、反対語のすくなくとも一つ以上の組合わせを用いることで主に類似性に関する情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【0019】請求項3、4、16、17、29及び30の発明は、各部分文書集合の代表語セットの関連語として反対語を用い、反対語が自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、その反対語を含む文書を文書集合から抽出し、それを新たな部分文書集合とすることで、文書集合からより多くの分析情報を抽出する文書分類装置、方法または記録媒体を提供することを目的とする。

【0020】また請求項5、18及び31の発明は、分類対象文書に形態素解析し、得られた解析結果をもとに分類対象文書を幾つかの文書集合に分類する文書分類装置において、形態素解析の結果得られる単語のうち指定

される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施すことによって、高品位な文書ベクトル空間を構成し、この文書ベクトル空間で統計処理を用いて文書分類を行うことで高品質な文書分類結果を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【0021】請求項6、19及び32の発明は、文書分類を行うための統計手法として、クラスタリング手法を用いることで、簡便に高品質な文書分類結果を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【0022】請求項7、20及び33の発明は、分類対象文書に形態素解析を適用することで抽出される単語の中で、特に、品詞が、接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、適切な結合処理を施すことで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【0023】請求項8、21及び34の発明は、単語の結合処理において特定の品詞の単語が出現するまで単語の結合を続けることによって新たな単語を生成することで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【0024】請求項9、22及び35の発明は、単語の結合処理において、品詞が数詞接尾詞もしくは助数詞の単語について、結合される複数の単語を削除し、文書ベクトル空間を生成する際にはそれらの単語の情報は用いないことで、高品質な文書ベクトル空間を得ることができる文書分類装置、方法または記録媒体を提供することを目的とする。

【0025】また本発明の請求項10～13、23～26及び36～39の発明では、事前に文書集合から様々な数の部分文書集合を生成し、生成された多数の部分文書集合間の関係を算出することで、ユーザが始めから文書集合の構造の把握を行う作業に集中できる情報を提供することを目的とする。

【0026】従って請求項10、23及び36の発明は、文書のベクトル空間モデルを用い、生成する部分文書集合の数をパラメータとして繰り返し文書分類処理を行うことで、多数の部分文書集合を生成し、さらに生成された多数の文書集合について相互の関係を算出することで、文書集合の構造の把握を支援する情報を生成する文書分類装置を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【0027】請求項11、24及び37の発明は、文書分類を行う統計手法として、非階層クラスタリング手法を用いることで、簡便に多数の部分文書集合を生成する文書分類装置、方法または記録媒体を提供することを目

的とする。

【0028】請求項12、25及び38の発明は、生成された多数の文書集合について相互の関係として、類似関係と包含関係を算出することで、容易に文書集合の構造を把握する情報を提供する文書分類装置、方法または記録媒体を提供することを目的とする。

【0029】請求項13、26及び39の発明は、生成された多数の文書集合が有する情報のうち、単語に関する情報のみを用いて相互の関係を算出することで、汎用性・再利用性の高い関係情報を算出する文書分類装置、方法または記録媒体を提供することを目的とする。

【0030】

【課題を解決するための手段】請求項1の発明は、文書集合をその内容に従って分類する文書分類装置であって、複数の文書を入力する文書入力部と、該文書入力部にて入力された各文書から該各文書を構成する単語情報を抽出する文書解析部と、該文書解析部にて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類部と、該文書分類部にて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出部と、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出部にて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出部と、該関連語抽出部にて抽出した関連語セットと前記代表語抽出部で抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成部と、前記文書分類部での分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存部とを含むことを特徴としたものである。

【0031】請求項2の発明は、請求項1の発明において、前記関連語抽出部にて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴としたものである。

【0032】請求項3の発明は、請求項1の発明において、前記関連語抽出部にて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴としたものである。

【0033】請求項4の発明は、請求項1の発明において、前記関連語抽出部にて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しな

い反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成部をさらに含むことを特徴としたものである。

【0034】請求項5の発明は、文書の内容に従って文書の分類を行う文書分類装置であって、文書データを入力する文書入力部と、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類部とを含み、前記文書解析部にて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴としたものである。

【0035】請求項6の発明は、請求項5の発明において、前記文書分類部において統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴としたものである。

【0036】請求項7の発明は、請求項5または6の発明において、前記文書解析部において品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うことを特徴としたものである。

【0037】請求項8の発明は、請求項5ないし7のいずれか1の発明において、前記文書解析部において特定の品詞の単語が出現するまで単語の結合を続けることを特徴としたものである。

【0038】請求項9の発明は、請求項5ないし8のいずれか1の発明において、前記文書解析部において品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類部では削除した単語の情報を含まないことを特徴としたものである。

【0039】請求項10の発明は、文書の内容に従って文書データ集合を分類する文書分類装置であって、文書データ集合を入力する文書入力部と、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析部と、該文書解析部にて抽出された文書データの解析結果を記憶する文書解析結果記憶部と、前記文書解析部にて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成部と、該文書ベクトル空間

生成部にて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶部と、指定される条件から文書データ集合の分類数を決定する分類数決定部と、前記文書ベクトル空間生成部にて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類部と、該文書分類部で生成された分類結果を記憶する分類結果記憶部と、前記分類数決定部から前記分類結果記憶部までの処理を繰り返し行うか否かの判定をおこなう繰り返し判定部と、前記文書ベクトルデータ記憶部と前記分類結果記憶部に記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出部と、該部分文書集合間関係算出部にて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶部とを含むことを特徴としたものである。

【0040】請求項11の発明は、請求項10の発明において、前記文書分類部にて用いられる統計手法が非階層クラスタリング手法であることを特徴としたものである。

【0041】請求項12の発明は、請求項10または11の発明において、前記部分文書集合間関係算出部にて算出される関係が、類似関係と包含関係であることを特徴としたものである。

【0042】請求項13の発明は、請求項12の発明において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出されることを特徴としたものである。

【0043】請求項14の発明は、文書集合をその内容に従って分類する文書分類方法であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含むことを特徴としたものである。

【0044】請求項15の発明は、請求項14の発明に

において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせであることを特徴としたものである。

【0045】請求項16の発明は、請求項14の発明において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴としたものである。

【0046】請求項17の発明は、請求項14の発明において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含むことを特徴としたものである。

【0047】請求項18の発明は、文書の内容に従って文書の分類を行う文書分類方法であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換えることを特徴としたものである。

【0048】請求項19の発明は、請求項18の発明において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行うことを特徴としたものである。

【0049】請求項20の発明は、請求項18または19の発明において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語および品詞の置き換えを行うこ

とを特徴としたものである。

【0050】請求項21の発明は、請求項18ないし20のいずれか1の発明において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続けることを特徴としたものである。

【0051】請求項22の発明は、請求項18ないし21のいずれか1の発明において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いないことを特徴としたものである。

【0052】請求項23の発明は、文書の内容に従って文書データ集合を分類する文書分類方法であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返す行か否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含むことを特徴としたものである。

【0053】請求項24の発明は、請求項23の発明において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法であることを特徴としたものである。

【0054】請求項25の発明は、請求項23または24の発明において、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係であることを特徴としたものである。

【0055】請求項26の発明は、請求項25の発明において、前記部分文書集合間の関係は各部分文書集合か

ら抽出される単語情報のみを用いて算出されることを特徴としたものである。

【0056】請求項27の発明は、文書集合をその内容に従って分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、複数の文書を入力する文書入力ステップと、該文書入力ステップにて入力された各文書から該各文書を構成する単語情報を抽出する文書解析ステップと、該文書解析ステップにて抽出された各文書の単語情報をもとに前記複数の文書による文書集合をいくつかの部分文書集合に分類する文書分類ステップと、該文書分類ステップにて分類された各部分文書集合からそれらの代表語セットを抽出する代表語抽出ステップと、任意の単語についてその関連語が記述された関連語辞書を用いて前記代表語抽出ステップにて抽出した各部分文書集合の代表語セットそれぞれについて関連語セットを抽出する関連語抽出ステップと、該関連語抽出ステップにて抽出した関連語セットと前記代表語抽出ステップで抽出した代表語セットと各部分文書集合に所属する文書に関する情報とをもとに個々の部分文書集合及び部分文書集合間の関連情報を生成する部分文書集合情報生成ステップと、前記文書分類ステップでの分類結果を前記部分文書集合情報生成部にて生成された情報と合わせて保存する分類結果保存ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0057】請求項28の発明は、請求項27に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが、同義語、類義語、反対語のうちの少なくとも一つ以上の組合わせである文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0058】請求項29の発明は、請求項27に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語セットが少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出された反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0059】請求項30の発明は、請求項27に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記関連語抽出ステップにて抽出される関連語が少なくとも反対語を含み、ある部分文書集合の代表語セットから抽出さ

れた反対語セットが、自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、該一致しない反対語セットと代表語セットから反対語セットに対応する代表語を除いた単語セットを含む文書を文書集合から抽出し、あらたな部分文書集合を生成する処理を全部分文書集合に対し再帰的に繰り返す反意部分文書集合生成ステップをさらに含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

10 【0060】請求項31の発明は、文書の内容に従って文書の分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データを入力する文書入力ステップと、前記文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析情報から文書データを多次元ベクトル空間で表現するための文書ベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データの分類を行う文書分類ステップとを含み、前記文書解析ステップにて抽出される特定の品詞を有する単語を、該特定の品詞の品詞情報に基づき、該特定の品詞の前後に抽出される一つ以上の単語と結合することにより生成される単語と置き換え、かつ該特定の品詞の品詞情報も適切に置き換える文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

30 【0061】請求項32の発明は、請求項31に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにおいて統計手法としてクラスタリング法を用いることで文書データの分類を行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0062】請求項33の発明は、請求項31または32に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、単語及び品詞の置き換えを行う文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

40 【0063】請求項34の発明は、請求項31ないし33のいずれか1に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて特定の品詞の単語が出現するまで単語の結合を続ける文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0064】請求項35の発明は、請求項31ないし34のいずれか1に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書解析ステップにおいて品詞が数詞接尾詞もしくは助数詞の単語について、該数詞接尾詞もしくは助数詞の単語に結合される複数の単語を削除し、前記文書分類ステップでは削除した単語の情報を用いない文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0065】請求項36の発明は、文書の内容に従って文書データ集合を分類する文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、文書データ集合を入力する文書入力ステップと、すべての文書データに形態素解析を適用し、前記文書データを構成する単語をそれらの品詞情報等とともに抽出する文書解析ステップと、該文書解析ステップにて抽出された文書データの解析結果を記憶する文書解析結果記憶ステップと、前記文書解析ステップにて抽出された文書データの解析情報から前記文書データを多次元ベクトル空間で表現するためのベクトル空間を生成する文書ベクトル空間生成ステップと、該文書ベクトル空間生成ステップにて生成された文書ベクトル空間の各文書データのベクトルデータを記憶する文書ベクトルデータ記憶ステップと、指定される条件から文書データ集合の分類数を決定する分類数決定ステップと、前記文書ベクトル空間生成ステップにて生成した文書ベクトル空間において統計手法を用いることにより文書データを前記指定された分類数の部分文書集合に分類する文書分類ステップと、該文書分類ステップで生成された分類結果を記憶する分類結果記憶ステップと、前記分類数決定ステップから前記分類結果記憶ステップまでの処理を繰り返し行うか否かの判定をおこなう繰り返し判定ステップと、前記文書ベクトルデータ記憶ステップと前記分類結果記憶ステップにて記憶された情報を用いて生成されたすべての部分文書集合間の関係情報を算出する部分文書集合間関係算出ステップと、該部分文書集合間関係算出ステップにて生成された部分文書集合間の関係情報を記憶する部分文書集合間関係記憶ステップとを含む文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0066】請求項37の発明は、請求項36に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記文書分類ステップにて用いられる統計手法が非階層クラスタリング手法である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0067】請求項38の発明は、請求項36または37に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体におい

て、前記部分文書集合間関係算出ステップにて算出される関係が、類似関係と包含関係である文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0068】請求項39の発明は、請求項38に記載の文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体において、前記部分文書集合間の関係は各部分文書集合から抽出される単語情報のみを用いて算出される文書分類方法を実行するためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0069】

【発明の実施の形態】本発明の実施例の説明においては、自然言語で記述された1つ以上の文の集まりで、それが分類対象となる場合は、これを文書と言う。また、ひとつの文書の終端には、それが判別可能な文書終端記号が付置されているものとする。具体的な例をあげれば、公開特許公報や特定の新聞記事も文書であるし、それらから請求項や特定の1文を取り出したものであってもこれを文書と見なす。

【0070】図1は本発明の請求項1, 2, 14, 15, 27及び28の発明に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部101は、キーボード、OCR装置、ハードディスク等の補助記憶装置等の入力手段が文書分類装置100に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。図2は、文書データを入力する処理の一例を示すフローチャートである。

【0071】図1における文書解析部102では、入力された文書それぞれに対し、自然言語解析を行い、単語やその品詞などを抽出する。さらに、文書内での単語の出現順序や、文書の作成者や作成日などの文書のメタ情報なども含めることができる。その後、文書群で出現した単語に対しユニークな単語IDを付与し、文書内での単語出現回数を計数する。一例として、文書に対し形態素解析を適用することで、文書内の単語表記と品詞を抽出し、その結果をもとに文書群で出現したユニークな単語の表記、品詞、識別番号を抽出し、また各文書を抽出されたユニークな単語識別番号とその頻度で表現する例を示すこととし、そのフローチャートを図3に示す。

【0072】例えば、図4(A)に示す文書1と文書2に対し、形態素解析を適用すると図4(B)のような結果が得られる。図4(B)において各切り出された単語の下の数値はそれらの品詞を示しており、その対応表は図4(C)に示す。文書群が図4(A)に示す2つの文書のみで構成されているとすると、文書群で出現したユニークな単語の表記、品詞、識別番号と各文書を単語識別番号とその頻度で表現した結果は図5(A)～図5(C)のようになる。ただし、簡単のため品詞としては

名詞と未登録語のみを採用する。

【0073】文書分類部103では、文書解析部102で生成された情報をもとに文書群の分類をおこなう。本発明では、分類手法は特に限定しないが、ここでは一例として、上記文書解析部102における実施例を継承して、各文書を文書群でユニークな単語の出現頻度のベクトルで表現し、これらのベクトルをもとにクラスタリング手法の1つである *k means* 法を用いて文書分類を行う例を示すこととし、そのフローチャートを図6に示す。ここで、ベクトル間の類似度は0-1の間の実数、かつ最大類似度は1であるとする。

【0074】図7(A)に示す15個の文書を図3及び図5に示すアルゴリズムを基に3つの部分文書集合に分類した結果を図7(B)に示す。ここで、品詞としては名詞と未登録語のみを採用し、また *k means* 法における類似測度は余弦測度であり、反復停止条件は繰返し回数5回としている。代表語抽出部104では、文書解析部102で生成した各文書の単語情報及び文書分類部で生成した部分文書グループに関する情報をもとに各部分文書集合における代表語セットを抽出する。

【0075】本発明では、代表語の抽出方法を特に限定しないが、ここでは一例として、上記文書分類部における実施例を継承し、各部分文書集合においてそれらに所属する文書をひとつの仮想的な文書とみなした時の、文書群でユニークな単語の出現頻度が指定されたしきい値以上の単語をそれらの部分文書集合の代表語セットとする例を示すこととし、そのフローチャートを図8に示す。

【0076】上記文書分類部における実施例の各部分文書集合について上記のフローチャートに従って求めた代表語セットを図9に示す。ここで、出現頻度のしきい値は2としている。

【0077】関連語抽出部105では、代表語抽出部104にて抽出した各部分文書集合の代表語それぞれについて、関連語辞書を用いて関連語を抽出し、それらを各部分文書集合の関連語セットとする。関連語辞書としては、同義語辞書、広義語辞書、狭義語辞書、類義語辞書、反対語辞書、兄弟語辞書、上位概念語辞書、下位概念語辞書等を用いることができるが、ここでは一例として、上記代表語抽出部における実施例を継承し、任意の一つの関連語辞書を用いて各部分文書集合の関連語セットを求める例を示すこととし、そのフローチャートを図10に示す。なお、複数の辞書を用いる場合には、前記処理を各辞書について繰返し行えばよい。簡単のため関連語として同義語のみを扱うとして、前記代表語抽出部の実施例で求めた各代表語の同義語が図11(A)に示されるような場合、各部分文書集合の関連語セットは図11(B)のように示される。

【0078】部分文書集合情報生成部106では、文書解析部102で生成した各文書の単語情報、文書分類部

103で生成した部分文書グループに関する情報、代表語抽出部104で抽出した各部分文書集合の代表語セット、及び関連語抽出部105で生成した各部分文書集合の関連語セットを基に個々の部分文書集合及び部分文書集合間の関連情報を生成する。

【0079】各部分文書集合固有の情報としては、代表語セットの集合、関連語セットの集合、及び各部分文書集合が多重分類を許す分類手法により生成されている場合は、代表語及び／または関連語を指定されるしきい値個数以上含む部分文書集合に所属する文書の部分集合等の情報を用いることができる。また、部分文書集合間の関連情報としては、部分文書集合間の代表語セット集合の積集合や和集合や差集合、関連語セットの集合の積集合や和集合や差集合、及び各部分文書集合が多重分類を許す分類手法により生成されている場合は、部分文書集合に所属する文書の積集合や和集合や差集合、代表語及び／または関連語を多く含む部分文書集合に所属する文書の部分集合間の積集合や和集合や差集合等の情報を用いることができる。

【0080】ここでは一例として、上記関連語抽出部における実施例を継承し、文書部分集合が多重分類を許す分類手法により生成されているとしたときに、部分文書集合情報として、代表語セットの集合、関連語セットの集合、部分文書集合間の代表語セット集合の積集合と和集合と差集合、部分文書集合間の関連語セット集合の積集合と和集合と差集合を生成する例を示すこととし、そのフローチャートを図12に示す。これらの情報により、特に部分文書集合間の類似性、関連性、及び包含関係などを把握することが可能になる。

【0081】分類結果保存部107では、文書解析部102で生成した各文書の単語情報、文書分類部103で生成した部分文書グループに関する情報、代表語抽出部104で抽出した各部分文書集合の代表語セット、関連語抽出部105で生成した各部分文書集合の関連語セット、及び部分文書集合情報生成部106で生成した個々の部分文書集合及び部分文書集合間の関連情報を適切な形式で保存する。保存された関連情報は、出力部108からユーザの要求に応じて、または予め定められた条件に従って所定の出力手段に適宜出力される。

【0082】図13は本発明の請求項3、4、16、17、29及び30に対応する実施例を説明するための文書分類装置200のブロック構成図である。なお、図1と同様の機能を有する部分には図1と同一の番号を付している。反意部分文書集合生成部201では、関連語抽出部105にて生成される関連語としてすくなくとも反対語が抽出されるとき、任意の文書部分集合が有する反対語が、自分を含む他のどの部分文書集合の代表語とも一致しない場合、この反対語を含む文書を文書群から抽出し、それを新しい部分文書集合とする処理をすべての部分文書集合について再帰的におこなう。

【0083】ここでは一例として、上記実施例を継承して、関連語抽出部105にて反対語のみが抽出されることとし、各部分文書集合が有する反意語セットについてそれが自分を含む他の部分文書集合の代表語と一致するか否かを判定し、反意語がどの代表語とも一致しない場合、検索手法を用いて文書群からその反意語を含む文書を抽出し、それらを新しい部分文書集合とする例を示すこととし、そのフローチャートを図14に示す。

【0084】例えば、図7(A)に示す文書群を分類した結果得られている図7(B)の部分文書集合3の代表語セットに着目してみる。この場合、代表語“商用”の反対語として、“無料、フリー”という単語が得られたとする。この場合、これらの単語はどの代表語とも一致せず、単語“無料”で文書群を検索した結果は該当0件であるが、単語“フリー”で検索した場合は、文書4、文書5、文書12が検索される。これをあらたな部分文書集合とした場合、代表語として、“リナックス、フリー、ディストリビューション”を得ることができる。

【0085】これにより文書群から任意の部分文書集合とは反対の意味を有する部分文書集合が文書分類部では生成されなかった場合にも、反対の意味を有する部分文書集合を生成することができるため、文書群からより広範囲な話題を抽出することが可能となる。

【0086】請求項4、17、30の発明では、反対語からあらたな部分文書集合を求める際に、反対語を生成した代表語以外の部分文書集合の代表語も合わせて部分文書集合を求めることにより、より対象の部分文書集合とは反対の意味をもつ部分文書集合を生成することが可能となるが、基本的な処理は上記実施例と同様の処理で求めることができる。すなわち、例えば、図14に示す

#### ○接頭詞全般

もし{対象単語の品詞が接頭詞である}ならば{

計数用変数：iに1を代入する

繰り返す{

対象単語の先頭に対象単語よりi回前に抽出された単語を結合させる

もし{i回前に抽出されている単語の品詞が分類時使用品詞である}ならば{

繰り返しループを抜ける

}

さもなくば{

iを1増加する

}

}

対象単語の品詞を変更する

}

【0091】

#### ○数詞接尾詞以外の接尾詞全般

もし{対象単語の品詞が数詞接尾詞以外の接尾詞である}ならば{

計数用変数：iに1を代入する

繰り返す{

るステップを反対語と反対語を生成した代表語以外の部分文書集合の代表語を組合わせた論理式を用いればよい。

【0087】図15は、本発明の請求項5～9、18～22及び31～35に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部301は、キーボード、OCR装置、ハードディスク等の補助記憶装置等の入力手段が文書分類装置300に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。この際、各文書データを一意に識別するために、例えばユニークな数などの、識別子を各文書に割り当てる。

【0088】文書解析部302では、入力された文書それぞれに対し形態素解析を適用し、各文書を構成する単語を品詞情報等とともに抽出する。この際、抽出した単語を識別するために、抽出した単語のうちユニークな表記を持つものについては、ユニークな識別子を付置しておく。さらに、形態素解析の結果得られる単語のうち指定される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施す。例として、品詞が接頭詞全般、接尾詞全般、及び助数詞である単語について前記の結合及び置き換え処理を行う動作を説明する。

【0089】まず、本例では、前記の結合および置き換え処理を品詞が、1. 接頭詞全般、2. 数詞接尾詞以外の接尾詞全般、3. 数詞接尾詞もしくは数助詞の場合別に以下のような規則でおこなうこととする。ただし、本発明における結合及び置き換え処理の規則はこれらに限定するものではない。

【0090】

27

28

```

対象単語の終端に対象単語より i 回後に抽出された単語を結合させる
もし {i 回前に抽出されている単語の品詞が分類時使用品詞である} ならば {
  繰り返しループを抜ける
}
さもなくば {
  i を 1 増加する
}
}
対象単語の品詞を変更する
}

```

【0092】

○数詞接尾詞もしくは助数詞

```

もし {対象単語の品詞が数詞接尾詞もしくは助数詞である} ならば {
  繰り返す {
    もし {対象単語の直前に抽出されている単語の品詞が数詞である} ならば {
      対象単語の先頭に対象単語の直前に抽出された単語を結合させる
      対象単語の i 回前に抽出された単語を削除する
    }
    さもなくば {
      繰り返しループを抜ける
    }
  }
  対象単語の品詞を変更する
}

```

【0093】図16に示す6つの文書データを分類対象文書データとし、この文書データに対して形態素解析を適用し、単語及びそれらの品詞を抽出したものを図17に示す。ただし、本発明では形態素解析系については特に規定しない。また、分類時使用品詞を普通名詞、サ変名詞、固有名詞、数詞、形容詞、接頭詞全般、接尾詞全般、助数詞とした場合の文書データの解析結果を図18に示す。

【0094】図18に示されている結果において、品詞が接頭詞全般、接尾詞、もしくは数助詞である単語に対し前記規則に従い、結合・置き換え処理を施した結果を図19に示す。例えば、文書1における「千葉【普通名詞】、氏【固有名詞接尾詞】」という文字列は、数詞接尾詞以外の接尾詞全般の規則を用いて、「千葉【普通名詞】、千葉氏【固有名詞】」という文字列になり、また「1【数詞】、9【数詞】、5【数詞】、0【数詞】、年【助数詞】」という文字列は、数詞接尾詞もしくは助数詞の規則を用いて、「1950年【普通名詞】」という文字列になる。

【0095】文書ベクトル空間生成部303では、前記文書解析部にて抽出された各文書データの単語情報をもとに文書データをベクトル表現するための空間を生成する。例として、前記文書解析部での例をもとに、文書データ全体でユニークな単語の頻度により文書ベクトル空間を生成することとする場合の各文書データのベクトル表現を生成する動作を説明する。ただし、本発明では、

ベクトル空間生成手法はこれに限定するものではなく、例えば、全単語の線形変換によりベクトル空間を生成することもできる。

【0096】図18及び図19に示す文書解析結果からユニークな単語を抽出し、各文書での該当単語の頻度を計数し、それらの結果を、単語を列方向に、文書データを行方向に付置することで、行列表現したものをそれぞれ図20と図21に示す。これら行列において、列ベクトルが各文書データのベクトルデータとなる。

【0097】文書分類部304では、前記文書ベクトル空間生成部にて生成された文書データベクトルを統計手法を用いることで幾つかの集合に分類する。出力部305では、文書分類部304で分類された文書データベクトルの集合をユーザの要求に応じてまたは予め定められた条件に従って所定の出力手段に適宜出力する。文書分類部304における統計処理は様々なものが利用可能であるが、請求項5の発明ではアルゴリズムの簡潔さやパラメータの有無等の理由からクラスタリング手法を用いることに限定している。例として、前記文書ベクトル空間生成部での例をもとに、クラスタリング手法を用いて文書ベクトル进行分类する動作を説明する。

【0098】ここでは、クラスタリング手法の1つである Ward 法を用いることとし、また類似測度は標準化ユークリッド距離測度を使用する。なお、クラスタリング手法に関しては、“多変量解析入門（森北出版）”に詳しい。図20及び図21に示されている文書データ

30

40

50

に対し、Warrd法を適用した結果を図22と図23に示す。ここで、図20は前記結合・置き換えの処理を適用した結果で文書ベクトル空間を構成したデータであり、図21は結合・置き換え処理を適用していない結果で文書ベクトル空間を構成したデータである。また、図22と図23の図中の数値は各クラスタ間の距離である。

【0099】図22及び図23の結果を比較した場合、文書4の位置の差異が非常に特徴的であり、結合・置き換えの処理を適用した場合は、文書4は文書2や文書5と類似していると判断され、結合・置き換えの処理を適用しない場合は、文書4は文書1や文書6と類似していると判断される。主観的な語彙の適合度などから判断して文書4は〔文書2、文書5〕の集合よりも〔文書1、文書3、文書6〕の集合に含まれる方が適切であると思われる。従って、この結果から、結合・置き換えの処理を適用することにより、より質の高い文書ベクトル空間を構成でき、この文書ベクトル空間で分類処理をおこなうことで、質の高い文書分類結果を得ることができる。

【0100】図24は本発明の請求項10～13、23～26及び36～39に対応する実施例を説明するための文書分類装置のブロック構成図である。文書入力部401は、キーボード、OCR装置、ハードディスク等の補助記憶装置による入力手段が文書分類装置400に直接に、または、ネットワーク経由で接続され、このような入力手段から文書や文書群を獲得し、文書データを入力するインターフェースである。この際、各文書データを一意に識別するために、例えばユニークな数などの、識別子を各文書に割り当てる。

【0101】文書解析部402では、入力された文書それぞれに対し形態素語解析を適用し、各文書を構成する単語を品詞情報等とともに抽出する。この際、抽出した単語を識別するために、抽出した単語のうちユニークな表記を持つものについては、前記文書データと同様にユニークな識別子を付置しておく。例として、文書データに対し形態素解析を適用し、文書データ全体で表記と品詞がユニークである単語を同定し、それらに一意な識別番号を付与するとともに、各文書データを、それを構成する単語の識別番号とその出現頻度を表現するための疑似コードを図25に示す。なお、本発明では、形態素解析系は必要な情報を抽出できるものであれば、どのようなものでよい。

【0102】文書解析結果記憶部403では、文書解析部402にて抽出された文書データの形態素解析結果を適切な形式で記憶する。文書ベクトル空間生成部404では、文書解析部402にて抽出された各文書データの単語情報をもとに文書データをベクトル表現するための空間を生成する。例として、文書解析部402での例をもとに、文書データ全体でユニークな単語の正規化された頻度により文書ベクトル空間を生成する場合の、各文

書データのベクトル表現を生成する疑似コードを図26に示す。ただし、本発明では、ベクトル空間生成手法はこれに限定するものではなく、例えば、特異値分解などを使用して全単語の線形変換によりベクトル空間を生成することもできる。

【0103】文書ベクトルデータ記憶部405では、文書ベクトル空間生成部404にて生成された文書データベクトルを適切な形式で記憶する。分類数決定部406では、繰り返し文書分類を行う際の分類数を決定する（分類数を定数×繰り返し数とした場合の疑似コードを図27に含む）。文書分類部407では、文書ベクトル空間生成部404にて生成された文書データベクトルを統計手法を用いることで分類数決定部集合に分類する。

【0104】統計処理は様々なものが利用可能であるが、請求項11の発明ではアルゴリズムの簡潔さやクラスタ数の変化により分類構造が動的に変化する特性等から非階層クラスタリング手法を用いることに限定している。例として、クラスタ数を繰り返し数と定数Nを乗じた数としてクラスタリング手法を用いて文書ベクトル进行分类する疑似コードを図27に示す。ここでは、クラスタリング手法の1つであるk means法を一部変更したものを用いることとし、また類似測度は余弦測度を使用する。なお、クラスタリング手法に関しては、“多変量解析入門（森北出版）”に詳しい。

【0105】文書分類結果記憶部408では、文書分類部407で生成される文書分類結果を適切な形式で記憶する。繰り返し判定部409では、繰り返し文書分類をおこなう際の繰り返しを継続するか否かの判定を行う

（繰り返し判定を指定された最大数を限度とした場合の疑似コードを図27に含む）。部分文書集合間関係算出部410では、文書分類結果記憶部408に記憶されている複数の部分文書集合間の関係情報を、文書解析結果記憶部403と文書ベクトルデータ記憶部405にて記憶されている種々の文書データに関する情報を用いて算出する。例として、部分文書集合間の類似関係と包含関係を文書データ及び／または文書データを構成する単語情報で算出する動作を説明する。

【0106】まず、部分文書集合間の類似関係と包含関係を文書データで表現するための定式化を行う。文書分類結果記憶部408に記憶されている複数の部分文書集合はユニークな識別番号が付与されているものとする。第m番目の部分文書集合の特性ベクトル：Vmを以下のように定義する。

【0107】・Vmの次元数は全文書データ数に等しい  
・Vmの各要素はそれぞれ1つの文書データに対応し、重複はない。

・要素iに対応する文書データと部分文書集合との類似度が閾値以上の場合、要素iは1となる。

・要素iに対応する文書データと部分文書集合との類似度が閾値未満の場合、要素iは0となる。

【0108】上記定義を用いて、第 $m$ 番目の部分文書集合と第 $n$ 番目の部分文書集合の関係： $R_{mn}$ と $R_{nm}$ を以下のように定義する。

$$(1) R_{mn} = \langle V_m, V_n \rangle / \langle V_m, V_m \rangle$$

$$(2) R_{nm} = \langle V_m, V_n \rangle / \langle V_n, V_n \rangle$$

ただし、 $\langle$ 、 $\rangle$ は内積を示す。

【0109】上記の $R_{mn}$ と $R_{nm}$ の値により、部分文書集合間の類似関係と包含関係を算出することが可能となる。図28は $R_{mn}$ と $R_{nm}$ の値による幾何学的解釈を示したものである。すなわち、 $R_{mn}$ が1に近い場合は、部分文書集合 $m$ は部分文書集合 $n$ に包含されているといえる。また、 $R_{mn}$ と $R_{nm}$ が両方1に近いほど部分文書集合 $m$ と部分文書集合 $n$ は類似しているものといえる。さらに、 $(R_{mn}, R_{nm})$ が $R_{mn} = R_{nm}$ の直線に近いほど、同じ程度の割合で相互に文書データを包含していることなども読み取れる。

【0110】次に、部分文書集合間の類似関係と包含関係を文書データを構成する単語の出現頻度情報で表現するための定式化をおこなう。第 $m$ 番目の部分文書集合の特性ベクトル： $W_m$ を以下のように定義する。

【0111】 $W_m$ の次元数は全文書データでユニークな単語数に等しい。 $W_m$ の各要素はそれぞれユニークな単語に対応し、重複はない $W_m$ の第 $i$ 番目の要素値を $w_m(i)$ と示す。部分文書集合との類似度が閾値以上の文書すべてにおける、要素 $i$ に対応する単語の出現頻度（出現回数）を要素 $i$ の要素値とする。

【0112】上記定義を用いて、第 $m$ 番目の部分文書集合と第 $n$ 番目の部分文書集合の関係： $R'_{mn}$ と $R'_{nm}$ を以下のように定義する。

$$(3) R'_{mn} = \sum f(w_m(k), w_n(k)) / \sum f(w_m(k), w_m(k))$$

$$(4) R'_{nm} = \sum f(w_m(k), w_n(k)) / \sum f(w_n(k), w_n(k))$$

$$(5) f(w_m(k), w_n(k)) = 0 \quad \text{for } w_m(k) \times w_n(k) = 0 \\ w_m(k) \times (a + b / |w_n(k) - w_m(k)| + 1) \quad \text{for } w_m(k) \times w_n(k) \neq 0$$

ただし、 $a, b$ は定数で、 $a + b = 1$ ,  $a, b \geq 0$

【0113】上記の $R'_{mn}$ と $R'_{nm}$ の値を用いても図28に示す $R_{mn}$ と $R_{nm}$ の関係と同様の解釈ができ、したがって、部分文書集合間の類似関係と包含関係を算出することが可能となる。さらに、 $R'_{mn}$ と $R'_{nm}$ を用いて部分文書集合の関係を定義する場合、文書データのレベルでは得ることのできない関係を得ることが可能になるとともに、例えば内容は一致してても、分析対象の文書データが異なっている場合にも部分文書集合間の関係を算出することが可能となる。また、部分文書集合間関係記憶部411では、部分文書集合間関係算出部410にて生成された部分文書集合間の関係情報を適切な形式で記憶する。また、出力部412は、部分文書集合間関係記憶部411で記憶された関係情報をユーザの要求に応じて、または予め定められた条件に従って出力手段に適宜出力する。

【0114】

【発明の効果】請求項1、14及び27の発明によれば、生成された部分文書集合それぞれの代表語セットを抽出し、さらにそれら代表語それぞれについて関連語を求め、これらの情報をもとに各部分文書集合および部分文書集合間の関連情報を生成することで、部分文書集合の分析に有効な情報を提供することができる。

【0115】請求項2、15及び28の発明によれば、関連語として同義語、類義語、反対語のすくなくとも一つ以上の組合わせを用いることで主に類似性に関する情報を提供することができる。

【0116】請求項3、4、16、17、29及び30の発明によれば、各部分文書集合の代表語セットの関連語として反対語を用い、反対語が自分を含む他のどの部分文書集合の代表語セットとも一致しない場合、その反対語を含む文書を文書集合から抽出し、それを新たな部分文書集合とすることで、文書集合からより多くの分析情報を抽出することができる。

【0117】請求項5、18及び31の発明によれば、分類対象文書に形態素解析し、得られた解析結果をもとに分類対象文書を幾つかの文書集合に分類する文書分類装置において、形態素解析の結果得られる単語のうち指定される品詞をもつ単語について、その前後の単語と適切に組合わせた単語と置き換え、かつ品詞もまた適切なものに置き換える処理を施すことによって、高品位な文書ベクトル空間を構成し、この文書ベクトル空間で統計処理を用いて文書分類を行うことで高品質な文書分類結果を得ることができる。

【0118】請求項6、19及び32の発明によれば、文書分類をおこなうための統計手法として、クラスタリング手法を用いることで、簡便に高品質な文書分類結果を得ることができる。

【0119】請求項7、20及び33の発明によれば、分類対象文書に形態素解析を適用することで抽出される単語の中で、特に、品詞が、接頭詞、接尾詞、助数詞、及びそれらに類する品詞である単語について、適切な結合処理を施すことで、高品質な文書ベクトル空間を得ることができる。

【0120】請求項8、21及び34の発明によれば、単語の結合処理において特定の品詞の単語が出現するまで単語の結合を続けることによって新たな単語を生成することで、高品質な文書ベクトル空間を得ることができる。

【0121】請求項9、22及び35の発明によれば、単語の結合処理において、品詞が数詞接尾詞もしくは助数詞の単語について、結合される複数の単語を削除し、文書ベクトル空間を生成する際にはそれらの単語の情報は用いないことで、高品質な文書ベクトル空間を得ることができる。

【0122】請求項10、23及び36の発明によれば

ば、文書のベクトル空間モデルを用い、生成する部分文書集合の数をパラメータとして繰り返し文書分類処理をおこなうことで、多数の部分文書集合を生成し、さらに生成された多数の文書集合について相互の関係を算出することで、文書集合の構造の把握を支援しうる情報を生成する文書分類装置を提供することができる。

【0123】請求項11、24及び37の発明によれば、上記目的に加え、文書分類をおこなう統計手法として、非階層クラスタリング手法を用いることで、簡便に多数の部分文書集合を生成することができる。

【0124】請求項12、25及び38の発明によれば、上記目的に加え、生成された多数の文書集合について相互の関係として、類似関係と包含関係を算出することで、容易に文書集合の構造の把握しうる情報を提供することができる。

【0125】請求項13、26及び39の発明によれば、上記目的に加え、生成された多数の文書集合が有する情報のうち、単語に関する情報のみを用いて相互の関係を算出することで、汎用性・再利用性の高い関係情報を算出することができる。

【図面の簡単な説明】

【図1】 本発明の請求項1、2、14、15、27及び28の発明に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図2】 文書データを入力する処理の一例を示すフローチャートである。

【図3】 文書に対し形態素解析を適用する処理の一例を示すフローチャートである。

【図4】 形態素解析の適用例について説明するための図である。

【図5】 形態素解析の適用結果の一例について説明するための図である。

【図6】 文書解析部で生成された情報をもとに文書群の分類を行う処理の一例を示すフローチャートである。

【図7】 文書の部分文書集合への分類を説明するための図である。

【図8】 代表語の抽出の処理の一例を示すフローチャートである。

【図9】 図8に示すフローチャートに従って求めた代表語セットの一例を示す図である。

【図10】 各部分文書集合の代表語セットのそれぞれについて関連語辞書を用いて関連語を抽出する処理の一例を示すフローチャートである。

【図11】 代表語抽出部で求め各代表語の同義語及び各部分文書集合の関連語セットの一例を示す図である。

【図12】 抽出または生成した代表語セット及び関連語セットを共に個々の部分文書集合及び部分文書集合間の関連情報を生成する処理の一例を示すフローチャート

である。

【図13】 本発明の請求項3、4、16、17、29及び30に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図14】 反対語を含む文書を抽出して新しい部分文書集合とする処理の一例を示すフローチャートである。

【図15】 本発明の請求項5～9、18～22及び31～35に対応する実施例を説明するための文書分類装置のブロック構成図である。

10 【図16】 分類対象文書データの例を示す図である。

【図17】 図16に示す文書データに形態素解析を適用して単語及び品詞を抽出した例を示す図である。

【図18】 文書データの解析結果の一例を示す図である。

【図19】 文書データの解析結果の他の例を示す図である。

【図20】 文書データを行方向に位置することで行列表現した例を示す図である。

20 【図21】 文書データを行方向に位置することで行列表現した他の例を示す図である。

【図22】 図20の文書データに対しWarrd法を適用した結果を示す図である。

【図23】 図21の文書データに対しWarrd法を適用した結果を示す図である。

【図24】 本発明の請求項10～13、23～26及び36～39に対応する実施例を説明するための文書分類装置のブロック構成図である。

【図25】 文書データの単語の識別番号とその出現頻度を表現するための擬似コードの一例を示す図である。

30 【図26】 各文書データのベクトル表現を生成する擬似コードの一例を示す図である。

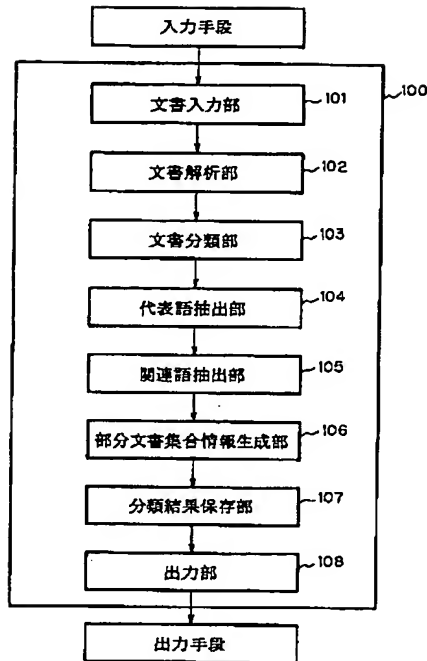
【図27】 クラスタリング手法を用いて文書ベクトルを分類する擬似コードの一例を示す図である。

【図28】 RmnとRnmの値による幾何学的解釈を示したものである。

【符号の説明】

100、200、300、400…文書分類装置、101、301、401…文書入力部、102、302、402…文書解析部、103、304、407…文書分類部、104…代表語抽出部、105…関連語抽出部、106…部分文書集合情報生成部、107…分類結果保存部、108、305、412…出力部、201…反意部分文書集合生成部、303、404…文書ベクトル空間生成部、403…文書解析結果記憶部、405…文書ベクトルデータ記憶部、406…分類数決定部、408…文書分類結果記憶部、409…繰り返し判定部、410…部分文書集合間関係算出部、411…部分文書集合間関係記憶部。

【図1】



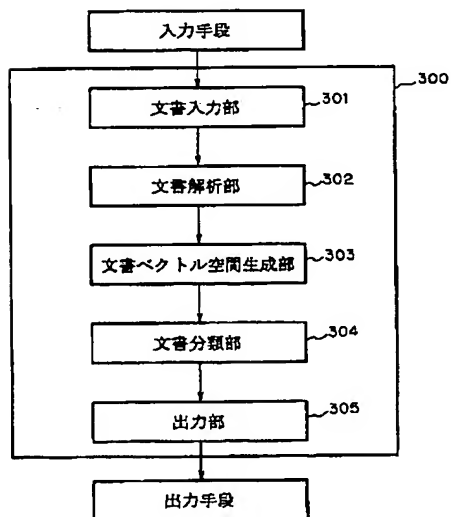
【図2】

文書群の文書数用変数:  $N$  を定義し、文書数を代入する。  
 大きさ  $N$  の文書配列変数:  $doc[n]$  ( $n=1?N$ ) を定義する。  
 繰り返し変数:  $i$  に 1 を代入する。  
 while ( $i$  が  $N$  以下である) {  
    $doc[i]$  に  $i$  番目の文書を代入する。  
    $i$  を 1 増加する。  
}

【図9】

| 部分文書集合 1  | 部分文書集合 2                                  | 部分文書集合 3                            | 部分文書集合 4 |
|---|---|-------------------------------------|----------|
| リナックス<br>カーネル<br>OS<br>可能<br>オペレーティングシステム<br>配布 | リナックス<br>日本語<br>環境<br>ディストリビューション<br>Vine | リナックス<br>商用<br>アプリケーション<br>開発<br>利用 | インストール   |

【図15】



【図16】

| 文書識別番号 | 文書内容                        |
|--------|-----------------------------|
| 1      | 千葉氏は1950年にフランス共和国に移住した。     |
| 2      | 近年フランスに移住を希望する人が増加している。     |
| 3      | 千葉商會にチーズを1950個発注した。         |
| 4      | フランス産の場合、希望小売価格は1950円である。   |
| 5      | 千葉産のチーズの価格はフランス産に比べると格段に安い。 |
| 6      | そして、1950年当時、まだ星を沢山見ることが出来た。 |

【図3】

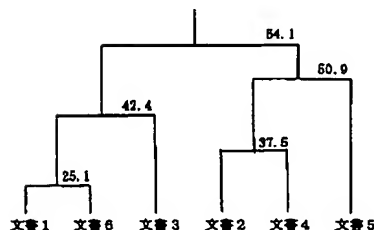
要素が大きさ可変長の単語識別番号保持用配列変数: iWordID [], 同じく可変長の大きさの単語頻度保持用配列変数: iWordFreq [], 単語識別番号数保持用変数: iWordSize, および文書識別番号保持用変数: iDocIDをもつ文書構成要素型: TYPE\_DocPartsを定義する。

```

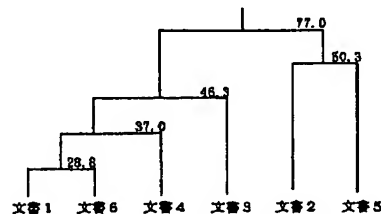
大きさ可変長の単語表配列変数: wordNote [] を定義する。
大きさ可変長の単語品詞用配列変数: wordClass [] を定義する。
大きさ可変長の単語識別番号用配列変数: wordID [] を定義する。
文書群のユニークな単語の単語数用変数: uWords を定義する。
大きさがNのTYPE_DocParts型配列変数: docParts [n] (n=1? N) を定義する。
繰り返し変数: iに1を代入する。
繰り返し変数: jに1を代入する。
uWordsに0を代入する。
while (iがN以下である) {
    docID[i] に形態素解析を適用する。
    抽出された単語数を変数: Mに代入する。
    繰り返し変数: jに1を代入する。
    docParts[i] のiDocIDにiを代入する。
    docParts[i] のiWordSizeにMを代入する。
    docParts[i] のiWordIDとiWordFreqの大きさをMとし、0を代入する。
    while (jがM以下である) {
        j番目に抽出された単語の表記を変数: wnに代入する。
        j番目に抽出された単語の品詞を変数: wcに代入する。
        繰り返し変数: kに1を代入する
        while (kがuWords以下である) {
            if (wnがwordNote[k] と一致かつwcがwordClass[k] と一致する) {
                繰り返しループから抜ける。
            }
            else {
                kを1増加する。
            }
        }
        // wnがすでに出現している単語の場合、kはuWords以下である
        if (kがuWordsより大きい) {
            uWordsを1増加する。
            wordNote、wordClass、wordIDを大きさを1大きくする。
            単語wnはユニークな単語であるので、wordNote[uWords] にwnを代入する。
            wordClassにwc[uWords] にwcを代入する。
            wordID[uWords] にuWordsを代入する。
        }
        docParts[i] のiWordID[j] にkを代入する。
    }
    docParts[i] のiWordIDを昇順にソートし、並び替えの順序情報を順序配列変数: mapに代入する。
    順序配列変数: mapをもとにdocParts[i] のiWordFreqを並び替える。
}

```

【図22】



【図23】



【図4】

(A)

文書1:

リナックス というのは自由に再配布することのできる、  
独立した Unix 系オペレーティングシステム(OS)のことです。

文書2:

リナックス は完全にスクラッチから書き起こされた OS であり、  
そのおかげであらゆる意味における「再配布」が可能なのです。

(B)

文書1:

リナックス という の は 自由 に 再 配 布 す る こ と の で き る 、

1        5 2    5 5 1    5 14 1    7    1    5 2        10

独立 し た Unix 系 オペレーティングシステム ( OS ) の こ と で す 、

1    7 7 8        14 1                                10 8    10 5    1    7    10

文書2:

リナックス は 完全 に スクラッチ から 書き 起こ さ れ た OS で あ り 、

1        5 1    5 8                                5 2    2        7 7 8 7 2    10

そ の お か げ で あ ら ゆ る 意 味 に お け る 「 再 配 布 」 が 可 能 な の で す 、

11 1        5 11        1    5 8        10 1 1    10 12 1    5 8        10

(C)

1 : 名詞、2 : 動詞

3 : 形容詞、4 : 形容動詞

5 : 助詞、6 : 副詞

7 : 助動詞、8 : 未登録語

9 : 数詞、10 : 記号

11 : 連体詞、12 : 接続詞

13 : 感動詞、14 : 接辞

【図5】

(A)

| 識別番号 (wordID) | 単語表記 (wordNote) | 品詞 (wordClass) |
|---------------|-----------------|----------------|
| 1             | リナックス           | 1              |
| 2             | 自由              | 1              |
| 3             | 再               | 1              |
| 4             | 配布              | 1              |
| 5             | 独立              | 1              |
| 6             | UNIX            | 8              |
| 7             | オペレーティングシステム    | 1              |
| 8             | OS              | 8              |
| 9             | 完全              | 1              |
| 10            | スクラッチ           | 8              |
| 11            | おかげ             | 1              |
| 12            | 意味              | 1              |
| 13            | おける             | 8              |
| 14            | 可能              | 1              |

文書1

(B)

| 単語識別番号<br>(IWordID) | 単語頻度<br>(iWordFreq) |
|---------------------|---------------------|
| 1                   | 1                   |
| 2                   | 1                   |
| 3                   | 1                   |
| 4                   | 1                   |
| 5                   | 1                   |
| 6                   | 1                   |
| 7                   | 1                   |
| 8                   | 1                   |

文書2

(C)

| 単語識別番号<br>(IWordID) | 単語頻度<br>(iWordFreq) |
|---------------------|---------------------|
| 1                   | 1                   |
| 9                   | 1                   |
| 10                  | 1                   |
| 8                   | 1                   |
| 11                  | 1                   |
| 12                  | 1                   |
| 13                  | 1                   |
| 3                   | 1                   |
| 4                   | 1                   |
| 14                  | 1                   |

【図6】

```

// 文書内のユニークな単語の頻度のベクトルの生成
文書内のユニークな単語の頻度のベクトル用の大きさがN×uWordsの2次元配列変数: DocVector
[N][uWords]を定義する。
繰り返し変数: iに1を代入する。
while (iがN以下である){
  繰り返し変数: jに1を代入する。
  繰り返し変数: kに1を代入する。
  while (jがuWords以下である){
    if (jがdocParts[i]のiWordID[k]と等しい){
      DocVector[i][j]にdocParts[i]のiWordFreq[k]を代入する。
      変数kを1増加する。
    }
    else {
      DocVector[i][j]を代入する。
    }
    変数jを1増加する。
  }
  変数iを1増加する。
}

// 書記クラス重心の設定
分類数用変数: Cを定義し、分類数を代入する。(CはN以下でなければならない)
クラス重心ベクトル用の大きさがC×uWordsの2次元配列変数: CVector[C][uWords]
を定義する。
繰り返し変数: iに1を代入する。
while (iがC以下である){
  繰り返し変数: jに1を代入する。
  while (jがuWords以下である){
    CVector[i][j]にDocVector[i][j]を代入する。
  }
}

// 精練化処理
// 多重分類の場合、DocPosの第2次元が1以上となる。
各文書の所属クラス番号用の大きさがN×可変である2次元配列変数: DocPos[N][]を定義する。こ
の場合、DocPos[N][1]である。
最大類似度用変数: maxSimを定義する。
類似度用変数: simを定義する。
while (反復停止条件を満たしていない){
  繰り返し変数: iに1を代入する。
  while (iがN以下である){
    繰り返し変数: jに1を代入する。
    maxSimに0を代入する。
    while (jがC以下である){
      i番目の文書: DocVector[i]とj番目のクラス重心CVector[j]との類似度を計算
      し、simに代入する。
      if (simがmaxSim以上である){
        DocPos[i][1]にjを代入する。
        maxSimにsimを代入する。
      }
    }
  }
  CVector[DocPos[i][1]]をCVector[DocPos[i]]とDocVector[i]
  の平均とする。
}
}

```

## 【図7】

## (A)

- 文書1: リナックス というのは自由に再配布することのできる、独立した Unix 系オペレーティングシステム(OS)のことです。
- 文書2: リナックス は完全にスクラッチから書き起こされた OS であり、そのおかげであらゆる意味における「再配布」が可能なのです。
- 文書3: リナックス は本来 OS の中核となるカーネル(kernel)だけを指す名称ですが、リナックス カーネルベースのシステム全体をさして「リナックス」と表現することもあります。
- 文書4: いくつかのディストリビューションのベンダーは、インストール用ソフトウェアを含めたディストリビューションを、インターネット上でフリーで入手できるようにしています。
- 文書5: リナックス カーネルのソースコードは常にフリーに入手可能でなくてはなりません。
- 文書6: リナックス がインターネットで自由に手に入るようになり、多くのベンダーは「ディストリビューション」と呼ばれるパッケージ化されたバージョンのリナックス を作りました。
- 文書7: リアルタイムなオペレーティングシステムを必要とする人々は、リナックス・カーネルを元にした小さなリアルタイム・カーネルに移行することを可能にしました。
- 文書8: 望めばリナックス に対して金銭をやり取りすることも可能ですが、その場合もリナックス の再配布を制限することはできません。
- 文書9: ソフトウェア開発、ネットワーク利用(インターネット、イントラネット、SOHO などなど)、そしてエンドユーザのプラットフォームとしてリナックス は高価な商用UNIX システムに取って代わりつつあるのです。
- 文書10: Vine リナックス は、日本語版のインストーラによって、特に初心者インストール時の不安を減らしています。
- 文書11: 海外の商用アプリケーションは RedHat リナックス を前提に開発されていることが多く、その多くをそのまま利用することができます。
- 文書12: 商用フォントやアプリケーションを含めた製品版が提供されても、Vine リナックス がフリーであることに変わりはありません。
- 文書13: 開発は、Slackware や RedHat リナックス に対応した日本語環境 add-on として定評のある、PJB のメンバーを中心に進められています。
- 文書14: Vine リナックス は、使いやすい日本語環境を提供する リナックス・ディストリビューションです。
- 文書15: インストールの直後から、快適な日本語環境で作業できるように、様々な配慮を行っております。

## (B)

| 部分文書集合1 | 部分文書集合2 | 部分文書集合3 | 部分文書集合4 |
|---------|---------|---------|---------|
| 文書7     | 文書14    | 文書12    | 文書15    |
| 文書5     | 文書13    | 文書11    | 文書4     |
| 文書3     | 文書10    | 文書9     |         |
| 文書1     | 文書8     |         |         |
| 文書2     | 文書6     |         |         |

【図8】

```

各部分文書集合における代表語セット用の大きさがC×可変の2次元配列変数: typicalWords [C]
[] を定義する。
各部分文書集合における代表語セット数用の大きさがCの配列変数: typicalWordsSize [C] を
定義する。
代表語判定をおこなうためのしきい値用変数: threshSize を定義し、指定された代表語判定をおこなう
ためのしきい値を代入する。
部分文書集合の単語頻度用の大きさがuWordsの配列変数: groupFreq [uWords] を定義する。
繰り返し変数: i に1を代入する。
while (iがC以下である) {
    groupFreqのすべての要素に0を代入する。
    繰り返し変数: j に1を代入する。
    // 部分文書集合の単語頻度を求める
    while (jがN以下である) {
        // 部分文書集合に所属する文書かどうかを判定する
        if (DocPos [j] [1] がiと一致する) {
            繰り返し変数: k に1を代入する。
            while (kがdocParts [j] [1] のiWordsSize以下である) {
                groupFreq [docParts [j] のiWordID [k]] にdocParts [j] の
                iWordFreq [k] を加える。
                kに1を加える。
            }
        }
        // 代表語セットの選定
        繰り返し変数: j に1を代入する。
        繰り返し変数: k に1を代入する。
        typicalWordsSize [i] に0を代入する。
        while (jはuWords以下である) {
            if (groupFreq [j] がthreshSize以上である) {
                typicalWords [i] [k] を確保する。
                typicalWords [i] [k] にwordNote [j] を代入する。
                typicalWordsSize [i] に1を加える。
                kに1を加える。
            }
        }
        jに1を加える。
    }
    iに1を加える。
}

```

【図19】

|         |      |       |      |       |      |
|---------|------|-------|------|-------|------|
| 文書1     |      | 文書2   |      | 文書3   |      |
| 千葉      | 普通名詞 | フランス  | 普通名詞 | 千葉    | 普通名詞 |
| 千葉氏     | 固有名詞 | 移住    | サ変名詞 | 商会    | 普通名詞 |
| 1950年   | 普通名詞 | 希望    | サ変名詞 | チーズ   | 普通名詞 |
| フランス    | 普通名詞 | 人     | 普通名詞 | 1950個 | 普通名詞 |
| フランス共和国 | 固有名詞 | 増加    | サ変名詞 | 発注    | サ変名詞 |
| 移住      | サ変名詞 |       |      |       |      |
| 文書4     |      | 文書5   |      | 文書6   |      |
| フランス    | 普通名詞 | 千葉    | 普通名詞 | 1950年 | 普通名詞 |
| フランス産   | 普通名詞 | 千葉産   | 普通名詞 | 量     | 普通名詞 |
| 希望      | サ変名詞 | チーズ   | 普通名詞 |       |      |
| 小売      | 普通名詞 | 価格    | 普通名詞 |       |      |
| 価格      | 普通名詞 | フランス  | 普通名詞 |       |      |
| 1950円   | 普通名詞 | フランス産 | 普通名詞 |       |      |
|         |      | 格段    | 普通名詞 |       |      |
|         |      | 安い    | 形容詞  |       |      |

【図10】

```

各部分文書集合における関連語セット用の大きさがC×可変の2次元配列変数: relativeWords[C]
[] を定義する。
各部分文書集合における関連語セット数用の大きさがCの配列変数: relativeWordsSize[C]
を定義する。
繰り返し変数: iに1を代入する。
総関連語数用変数: totalSizeを定義する。
while (iがC以下である) {
  繰り返し変数: jに1を代入する。
  totalSizeに0を代入する。
  while (jがtypicalWordSize[j]以下である) {
    if (typicalWord[j]に関連語が登録されている) {
      関連語辞書からtypicalWord[j]の関連語数を取得し、relativeWordsSize
      [i]に代入する。
      繰り返し変数: kに1を代入する。
      while (kがrelativeWordsSize[i]以下である) {
        関連語辞書から次の関連語を取得する。
        // 同じ表記の単語がすでに関連語として取得されているか否かのチェック
        繰り返し変数: lに1を代入する。
        while (lがtotalSize+k?1以下である) {
          if (取得した単語がrelativeWords[i][l]と一致する) {
            繰り返しループから抜ける。
          }
        }
        if (lがtotalSize+k?1より大きい) {
          // 新しい関連語の登録
          relativeWords[i][totalSize+k]を確保する。
          relativeWords[i][totalSize+k]に取得した単語を代入する。
          kに1を加える。
        }
        else {
          // すでにある単語は登録しない
          relativeWordsSize[i]から1を引く。
        }
      }
      totalSizeにrelativeWordsSize[i]を加える。
    }
    jに1を加える。
  }
  iに1を加える。
}

```

【図25】

```

文書データすべてに対し繰り返す {
  文書データに形態素解析を適用する
  抽出した単語すべてに対し繰り返す {
    もし抽出した単語がすでに文書データ構成単語リストに存在するならば {
      文書データ構成単語リストの該当単語の頻度に1を加える
    }
    さもなくば {
      もし抽出した単語がユニーク単語リストに存在しないならば {
        ユニーク単語リストに該当単語の表記と品詞を登録し、ユニークな識別番号を付与する
      }
      文書データ構成単語リストに該当単語の識別番号を追加し、その頻度を1にする
    }
  }
}

```

【図11】

(A)

| 単語           | 関連語数 | 関連語                           |
|--------------|------|-------------------------------|
| リナックス        | 0    | 可能性、可                         |
| カーネル         | 0    |                               |
| OS           | 0    |                               |
| 可能           | 2    |                               |
| オペレーティングシステム | 0    | 分配、分与、配給、配布、配達、配当、ディストリビューション |
| 配布           | 7    | 日語、ジャパニーズ、国語、母国語              |
| 日本語          | 4    | 雰囲気、アンビエンス                    |
| 環境           | 2    | 分配、配布、販売、流通                   |
| ディストリビューション  | 4    | アプリ                           |
| Vine         | 0    |                               |
| 商用           | 0    |                               |
| アプリケーション     | 1    |                               |
| 開発           | 7    | 開墾、開発、新開、展開、打開、突破、打破          |
| 利用           | 7    | 使用、行使、運用、活用、駆使、転用、採用          |
| インストール       | 1    | 設置                            |

(B)

| 部分文書集合 1  | 部分文書集合 2   | 部分文書集合 3  | 部分文書集合 4 |
|---|--|---|----------|
| 可能性<br>可<br>分配<br>分与<br>配給<br>配布<br>配達<br>配当<br>ディストリビューション | 日語<br>ジャパニーズ<br>国語<br>母国語<br>雰囲気<br>アンビエンス<br>分配<br>配布<br>販売<br>流通 | アプリ<br>開墾<br>開発<br>新開<br>展開<br>打開<br>突破<br>打破<br>使用<br>行使<br>運用<br>活用<br>駆使<br>転用<br>採用 | 設置       |

【図12】

```

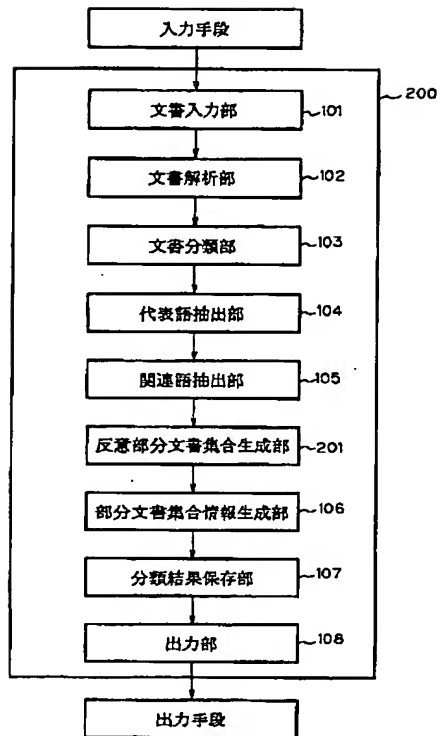
各部分文書集合における代表語セットは typicalWords [C] [] に保持されている。
各部分文書集合における関連語セットは relativeWords [C] [] に保持されている。
部分文書集合間の代表語セットの和集合用の2次元配列変数: typicalUnion [C×(C?1)/2] []
部分文書集合間の代表語セットの積集合用の2次元配列変数: typicalIntersection [C×(C?1)/2] []
部分文書集合間の代表語セットの差集合用の3次元配列変数: typicalDiff [C] [C] []
部分文書集合間の関連語セットの和集合用の2次元配列変数: relativeUnion [C×(C?1)/2] []
部分文書集合間の関連語セットの積集合用の2次元配列変数: relativeIntersection [C×(C?1)/2] []
部分文書集合間の関連語セットの差集合用の3次元配列変数: relativeDiff [[C] [C] []

// 部分文書集合間の代表語セットと関連語セットの和集合、積集合、差集合の生成
繰り返し変数: i に1を代入する。
繰り返し変数: j に1を代入する。
while (iがC以下である) {
  繰り返し変数: k に i+1 を代入する。
  while (kがC以下である) {
    // 部分文書集合 i と部分文書集合 k の代表語セットの和集合の取得
    typicalWords [i] と typicalWords [k] の和集合を typicalUnion [j]
    に代入する。
    // 部分文書集合 i と部分文書集合 k の代表語セットの積集合の取得
    typicalWords [i] と typicalWords [k] の積集合を
    typicalIntersection [j] に代入する。
    // 部分文書集合 i と部分文書集合 k の関連語セットの和集合の取得
    relativeWords [i] と relativeWords [k] の和集合を
    relativeUnion [j] に代入する。
    // 部分文書集合 i と部分文書集合 k の関連語セットの積集合の取得
    relativeWords [i] と relativeWords [k] の積集合を
    relativeIntersection [j] に代入する。
    j に1を加える。
  }
}

繰り返し変数: i に1を代入する。
while (iがC以下である) {
  繰り返し変数: j に1を代入する。
  while (jがC以下である) {
    // 部分文書集合 i と部分文書集合 j との代表語セットの差集合の取得
    typicalWords [i] から typicalWords [j] との共通要素を除いた集合を
    typicalDiff [i] [j] に代入する。
    // 部分文書集合 i と部分文書集合 j との関連語セットの差集合の取得
    relativeWords [i] から relativeWords [j] との共通要素を除いた集合を
    relativeDiff [i] [j] に代入する。
  }
}

```

【図13】



【図20】

|      | 文書 1 | 文書 2 | 文書 3 | 文書 4 | 文書 5 | 文書 6 |
|------|------|------|------|------|------|------|
| 千葉   | 1    | 0    | 1    | 0    | 1    | 0    |
| 氏    | 1    | 0    | 0    | 0    | 0    | 0    |
| 1    | 1    | 0    | 1    | 1    | 0    | 1    |
| 9    | 1    | 0    | 1    | 1    | 0    | 1    |
| 5    | 1    | 0    | 1    | 1    | 0    | 1    |
| 0    | 1    | 0    | 1    | 1    | 0    | 1    |
| 年    | 1    | 0    | 0    | 0    | 0    | 1    |
| フランス | 1    | 1    | 0    | 1    | 1    | 0    |
| 共和国  | 1    | 0    | 0    | 0    | 0    | 0    |
| 移住   | 1    | 1    | 0    | 0    | 0    | 0    |
| 希望   | 0    | 1    | 0    | 0    | 0    | 0    |
| 人    | 0    | 1    | 0    | 0    | 0    | 0    |
| 増加   | 0    | 1    | 0    | 0    | 0    | 0    |
| 商会   | 0    | 0    | 1    | 0    | 0    | 0    |
| チーズ  | 0    | 0    | 1    | 0    | 1    | 0    |
| 個    | 0    | 0    | 1    | 0    | 0    | 0    |
| 発注   | 0    | 0    | 1    | 0    | 0    | 0    |
| 産    | 0    | 0    | 0    | 1    | 2    | 0    |
| 小売   | 0    | 0    | 0    | 1    | 0    | 0    |
| 価格   | 0    | 0    | 0    | 1    | 1    | 0    |
| 円    | 0    | 0    | 0    | 1    | 0    | 0    |
| 格段   | 0    | 0    | 0    | 0    | 1    | 0    |
| 安い   | 0    | 0    | 0    | 0    | 1    | 0    |
| 星    | 0    | 0    | 0    | 0    | 0    | 1    |

【図18】

|   |   |   |
|---|---|---|
| <b>文書 1</b><br>千葉 普通名詞<br>氏 固有名詞接尾辞<br>1 数詞<br>9 数詞<br>5 数詞<br>0 数詞<br>年 助数詞<br>フランス 普通名詞<br>共和国 固有名詞接尾辞<br>移住 サ変名詞 | <b>文書 2</b><br>フランス 普通名詞<br>移住 サ変名詞<br>希望 サ変名詞<br>人 普通名詞<br>増加 サ変名詞                               | <b>文書 3</b><br>千葉 普通名詞<br>商会 普通名詞<br>チーズ 普通名詞<br>1 数詞<br>9 数詞<br>5 数詞<br>0 数詞<br>個 助数詞<br>発注 サ変名詞 |
| <b>文書 4</b><br>フランス 普通名詞<br>産 接尾辞<br>希望 サ変名詞<br>小売 普通名詞<br>価格 普通名詞<br>1 数詞<br>9 数詞<br>5 数詞<br>0 数詞<br>円 助数詞         | <b>文書 5</b><br>千葉 普通名詞<br>産 接尾辞<br>チーズ 普通名詞<br>価格 普通名詞<br>フランス 普通名詞<br>産 接尾辞<br>格段 普通名詞<br>安い 形容詞 | <b>文書 6</b><br>1 数詞<br>9 数詞<br>5 数詞<br>0 数詞<br>年 助数詞<br>星 普通名詞                                    |

【図14】

各部分文書集合における代表語セットは `typicalWords [C] []` に保持されている。  
 各部分文書集合における代表語セットの大きさは `typicalWordsSize [C]` に保持されている。  
 各部分文書集合における反対語（関連語）セットは `relativeWords [C] []` に保持されている。  
 各部分文書集合における反対語セットの大きさは `relativeWordsSize [C]` に保持されている。

部分文書集合数用変数: `groupSize` を定義し、`C` を代入する。  
 // `i` は反対語を有する部分文書集合  
 繰り返し変数: `i` に1を代入する。  
 while (`i` が `groupSize` 以下である) {  
 // `j` は各反対語  
 繰り返し変数: `j` に1を代入する。  
 while (`j` は `relativeWordsSize [i]` 以下である) {  
 // `k` は比較対象部分文書集合  
 繰り返し変数: `k` に1を代入する。  
 while (`k` は `groupSize` 以下である) {  
 // `l` は比較対象部分文書集合の代表語  
 繰り返し変数: `l` に1を代入する。  
 while (`l` は `typicalWordsSize [k]` 以下である) {  
 if (`relativeWords [i] [j]` が `typicalWords [k] [l]` と一致する) {  
 繰り返し変数: `k` のループの外側に抜ける。  
 }  
`l` に1を加える。  
 }  
`k` に1を加える。  
 }  
 // `k` が `groupSize` よりおおきければ、対象の反対語はどの関連語とも一致していない。  
 if (`k` が `groupSize` より大きい) {  
 検索手法を用い、文書群から `relativeWords [i] [j]` を含む文書を収集する。  
`groupSize` に1を加える。  
 繰り返し変数: `k` に1を代入する。  
 while (`k` が `N` 以下である) {  
 if (文書 `k` が検索された) {  
 対象となる `DocPos [k] []` に新しい要素を確保し、`groupSize` を代入する。  
 }  
 }  
`typicalWords`、`typicalWordsSize`、`relativeWords`、  
`relativeWordsSize` にそれぞれ要素の一つを追加し、部分文書集合 `groupSize` の代  
 表語セット、関連語セットを求める。  
 }  
`j` に1を加える。  
 }  
`i` に1を加える。  
 }

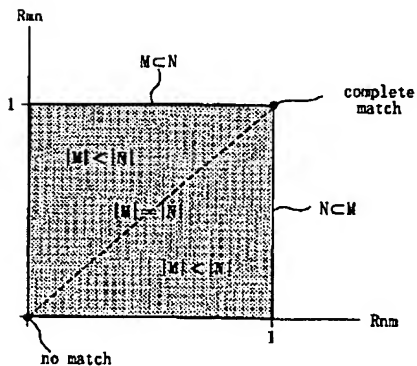
【図26】

文書データすべてに対し繰り返す {  
 次元数が文書データ全体でユニークな単語数の文書データベクトルを生成する  
 文書データベクトルの全要素に0を代入する  
 文書データ構成単語リストのすべての単語に対し繰り返す {  
 保持する識別番号に対応する文書データベクトルの要素に保持する単語頻度を代入する  
 }  
 文書データベクトルを長さ1になるように正規化する  
 }

【図17】

|  |   |   |
|--|---|---|
| <p>文書1</p> <p>千葉 普通名詞<br/>氏 固有名詞接尾辞<br/>は 副助詞<br/>1 数詞<br/>9 数詞<br/>5 数詞<br/>0 数詞<br/>年 助数詞<br/>に 格助詞<br/>フランス 普通名詞<br/>共和国 固有名詞接尾辞<br/>に 格助詞<br/>移住 サ変名詞<br/>した サ変助動詞<br/>た 助動詞<br/>。 句点記号</p>    | <p>文書2</p> <p>近年 副詞的名詞<br/>フランス 普通名詞<br/>に 格助詞<br/>移住 サ変名詞<br/>を 格助詞<br/>希望 サ変名詞<br/>する サ変助動詞<br/>人 普通名詞<br/>が 格助詞<br/>増加 サ変名詞<br/>し サ変助動詞<br/>て 接続助詞<br/>いる 動詞<br/>。 句点記号</p>                | <p>文書3</p> <p>千葉 普通名詞<br/>商会 普通名詞<br/>に 格助詞<br/>チーズ 普通名詞<br/>を 格助詞<br/>1 数詞<br/>9 数詞<br/>5 数詞<br/>0 数詞<br/>個 助数詞<br/>発注 サ変名詞<br/>した サ変助動詞<br/>た 助動詞<br/>。 句点記号</p>  |
| <p>文書4</p> <p>フランス 普通名詞<br/>産 接尾辞<br/>の 格助詞<br/>場合 副詞的名詞<br/>、 読点記号<br/>希望 サ変名詞<br/>小売 普通名詞<br/>価格 普通名詞<br/>は 格助詞<br/>1 数詞<br/>9 数詞<br/>5 数詞<br/>0 数詞<br/>円 助数詞<br/>で 格助詞<br/>ある 動詞<br/>。 句点記号</p> | <p>文書5</p> <p>千葉 普通名詞<br/>産 接尾辞<br/>の 格助詞<br/>チーズ 普通名詞<br/>の 格助詞<br/>価格 普通名詞<br/>は 格助詞<br/>フランス 普通名詞<br/>産 接尾辞<br/>に 格助詞<br/>比べる 動詞<br/>と 格助詞<br/>格段 普通名詞<br/>に 格助詞<br/>安い 形容詞<br/>。 句点記号</p> | <p>文書6</p> <p>そして 接続詞<br/>、 読点記号<br/>1 数詞<br/>9 数詞<br/>5 数詞<br/>0 数詞<br/>年 助数詞<br/>当時 副詞的名詞<br/>、 読点記号<br/>まだ 副詞<br/>屋 普通名詞<br/>を 格助詞<br/>沢山 副詞的名詞<br/>見る 動詞<br/>こと 形式名詞<br/>が 格助詞<br/>出来 動詞<br/>た 助動詞<br/>。 句点記号</p> |

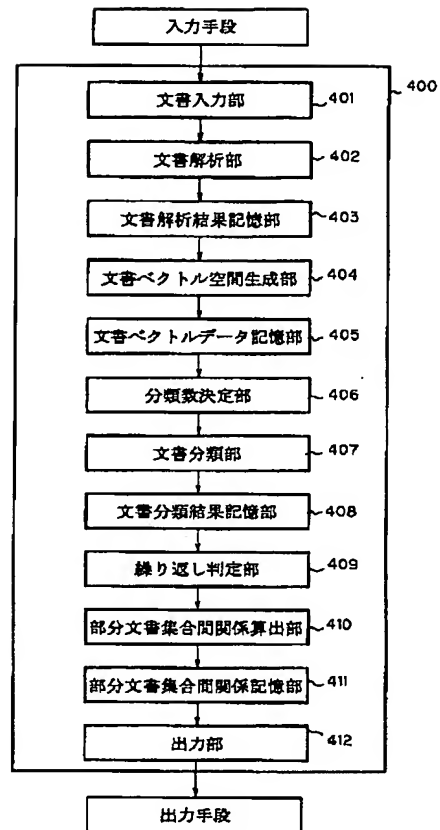
【図28】



【図21】

|         | 文書1 | 文書2 | 文書3 | 文書4 | 文書5 | 文書6 |
|---------|-----|-----|-----|-----|-----|-----|
| 千葉      | 1   | 0   | 1   | 0   | 1   | 0   |
| 千葉氏     | 1   | 0   | 0   | 0   | 0   | 0   |
| 1950年   | 1   | 0   | 0   | 0   | 0   | 1   |
| フランス    | 1   | 1   | 0   | 1   | 1   | 0   |
| フランス共和国 | 1   | 0   | 0   | 0   | 0   | 0   |
| 移住      | 0   | 1   | 0   | 0   | 0   | 0   |
| 希望      | 0   | 1   | 0   | 1   | 0   | 0   |
| 人       | 0   | 1   | 0   | 0   | 0   | 0   |
| 増加      | 0   | 1   | 0   | 0   | 0   | 0   |
| 商会      | 0   | 0   | 1   | 0   | 0   | 0   |
| チーズ     | 0   | 0   | 1   | 0   | 1   | 0   |
| 1950個   | 0   | 0   | 1   | 0   | 0   | 0   |
| 発注      | 0   | 0   | 1   | 0   | 0   | 0   |
| フランス産   | 0   | 0   | 0   | 1   | 1   | 0   |
| 小売      | 0   | 0   | 0   | 1   | 0   | 0   |
| 価格      | 0   | 0   | 0   | 1   | 1   | 0   |
| 1950円   | 0   | 0   | 0   | 1   | 0   | 0   |
| 千葉産     | 0   | 0   | 0   | 0   | 1   | 0   |
| 格段      | 0   | 0   | 0   | 0   | 1   | 0   |
| 安い      | 0   | 0   | 0   | 0   | 1   | 0   |
| 星       | 0   | 0   | 0   | 0   | 0   | 1   |

【図24】



【図27】

```
分類繰返し数: Lを1にする
最大繰返し数をMにする
繰返す {

  文書データベクトルからL×N個のベクトルをランダムに選出し、これらを重心ベクトルとする
  繰返す {
    すべての文書データベクトルに対し繰返す {
      対象文書データベクトルと最も余弦測度が1に近い重心ベクトルを求める
      上記選出された重心ベクトルを、それと上記対象文書データベクトルとの平均に置き換える
    }
    もし、繰返し数が許容値より大きい、重心と文書データの二乗平均誤差が許容値以下ならば {
      繰返しを抜ける
    }
  }

  すべての重心ベクトルに対し繰返す {
    対象重心ベクトルに固有の所属文書データリストを生成する
  }

  すべての文書データベクトルに対し繰返す {
    対象文書データベクトルと最も余弦測度が1に近い重心ベクトルを求め、
    該当重心ベクトルに固有の所属文書データリストに文書データベクトルの識別番号を加える
  }

  Lに1を加える
  もし、LがMより大きいならば {
    繰返しを抜ける
  }
}
```